BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs

Year: 2023 | Citations: 399 | Authors: Sheng Zhang, Yanbo Xu, N. Usuyama, J. Bagga, Robert Tinn

---

Abstract

Biomedical data is inherently multimodal, comprising physical measurements and natural language narratives. A generalist biomedical AI model needs to simultaneously process different modalities of data, including text and images. Therefore, training an effective generalist biomedical model requires high-quality multimodal data, such as parallel image-text pairs. Here, we present PMC-15M, a novel dataset that is two orders of magnitude larger than existing biomedical multimodal datasets such as MIMIC-CXR, and spans a diverse range of biomedical image types. PMC-15M contains 15 million biomedical image-text pairs collected from 4.4 million scientific articles. Based on PMC-15M, we have pretrained BiomedCLIP, a multimodal foundation model, with domain-specific adaptations tailored to biomedical vision-language processing. We conducted extensive experiments and ablation studies on standard biomedical imaging tasks from retrieval to classification to visual question-answering (VQA). BiomedCLIP achieved new state-of-the-art results in a wide range of standard datasets, substantially outperforming prior approaches. Intriguingly, by large-scale pretraining on diverse biomedical image types, BiomedCLIP even outperforms state-of-the-art radiology-specific models such as BioViL in radiology-specific tasks such as RSNA pneumonia detection. In summary, BiomedCLIP is a fully open-access foundation model that achieves state-of-the-art performance on various biomedical tasks, paving the way for transformative multimodal biomedical discovery and applications. We release our models at https://aka.ms/biomedclip to facilitate future research in multimodal biomedical AI.